Journal of Statistical Software

January 2010, Volume 32, Book Review 5. http://www.jstatsoft.org/

Reviewer: Qing Zhou
University of California, Los Angeles

A Guide to QTL Mapping with R/qtl

Karl W. Broman and Śaunak Sen
Springer-Verlag, New York, 2009.
ISBN 987-0-387-92124-2. 396 pp. USD 89.95.
http://www.Rqtl.org/book/

A quantitative trait, such as blood pressure and body weight, is a phenotypic characteristic that may be attributed to genetic and environmental factors. Quantitative trait loci (QTL) refer to DNA sites, often genes, that contribute to variation in a quantitative trait. QTL mapping is the detection of QTL in a population or an experimental cross. The book by Broman and Sen gives a practical review of statistical QTL mapping in experimental crosses with step-by-step instructions to the use of the R package **qtl**. The book has a wide coverage of topics, from experimental design and data input, to single-QTL mapping, mapping with covariates and multiple-QTL scans, organized into 11 chapters according to the logic flow of a QTL analysis. The authors made a good compromise between statistical methodology and real-data illustration with R code, so that different readers may easily focus on the parts that are more interesting to them. A researcher can follow the code in the examples of the book to study real-data applications of the **qtl** package.

The first chapter contains a brief introduction to experimental crosses, the underlying statistical problems, the software R, and the package **qtl**. The concepts of backcross and intercross experiments in genetics are discussed in the context of phenotype data analysis. This quickly builds a foundation to follow the main idea of the book even for a statistician with a limited background in genetics. The statistical structure of QTL mapping is presented with a diagram of four key components: QTL, genetic markers, phenotypes, and covariates. Genotype data at genetic markers provide the source information about QTL which influence phenotypes with a collection of covariates. This structure splits the QTL mapping problem into two familiar statistical problems: the missing data problem and the model selection problem. Genotype information at a QTL is often missing and thus needs to be inferred from nearby observed genetic markers. This is the missing data part of the problem. There are a large number of candidate QTL, and one needs to identify a small subset of them which actually affect a phenotype. This is a typical model selection problem with small n (number of individuals in a study) and large p (number of candidate QTL).

After an explanation of the input of QTL mapping data into R/**qtl** in Chapter 2 and a description of various data quality checks in Chapter 3, the book arrives at its core in Chapter 4,

single-QTL analysis. One assumes the presence of a single QTL and considers individually each site across the genome for the putative QTL. The authors begin with marker regression in which a proper test is performed at each genetic marker on the association between the marker and the phenotype. The focus of the chapter is the more sophisticated interval mapping approach, which differs from marker regression in that missing genotypes between observed markers are taken into account. The book discusses four different methods for interval mapping, the standard method with EM-based estimation of genotypes, Haley-Knott regression, extended Haley-Knott regression, and multiple imputation. Relative advantages and disadvantages of the four methods are summarized. The chapter also introduces methods to determine the statistical significance of a detected QTL under the global null hypothesis of no QTL. In addition, technical difficulties in analyzing X chromosome and interval estimation of the location of a QTL are mentioned. The conditional distribution of a phenotype given the genotypes of its QTL is assumed to be normal for all the methods in Chapter 4. This assumption is relaxed in Chapter 5 to consider non-normal phenotypes; in particular, the authors focus on rank-based nonparametric methods for interval mapping and a specific model for binary traits.

Starting from Chapter 6 the book moves on to more complicated and realistic problems in QTL analyses. To perform a solid genetic study on QTL mapping, a good experimental design is necessary. An experimenter needs to determine the type of cross, the density of genotyping, what covariates to include, and what phenotypes to measure. The authors discuss these issues in Chapter 6 and illustrate the use of the package **qtlDesign**. The topic of Chapter 7 is about the inclusion of covariates in QTL mapping. This is easily understood as a linear model, in which a QTL and covariates, such as gender, diet, and other environmental factors, are taken as predictors. The authors generalize the linear regression framework to models with non-normal phenotypes and with interaction terms between QTL and covariates. So far the book deals with single-QTL mapping. However, most complex traits are affected by multiple genetic loci. Serving as a transition from single-QTL detection to multiple-QTL methods, Chapter 8 describes two-QTL scan methods, emphasizing the detection of a second QTL by likelihood ratio-based tests.

In Chapter 9, the authors formulate multiple-QTL mapping as a model selection problem, and discuss four aspects of the problem: class of models, model fitting, model search, and model comparison. The class of models considered include additive QTL models and models with pairwise interactions between QTL. These models may be fitted with EM algorithms if the genotype data are regarded as missing. A forward/backward algorithm for model search is implemented in **qtl**, which is able to identify additive QTL and their two-way interactions. Model comparison is performed based on penalized likelihood (LOD score). The authors recommend an interesting principle to choose the penalty parameter such that the false positive inclusion rate is controlled at a desired level, say 5%. Then, they review briefly Bayesian QTL mapping as an alternative to the classical model selection approach, and compare the relative advantages of the two approaches. This chapter also contains an overview of available functions in **qtl** for multiple-QTL mapping, together with a detailed illustration of their use. The book concludes with two case studies which illustrate a complete QTL mapping process in the last two chapters.

In sum, this is a well-written book with carefully chosen and nicely organized topics. It can serve as a good introduction to QTL mapping methodology and a useful practical guide to the R package. A reader with an intermediate background in genetics and statistics will find no

difficulty in understanding the book or following the demonstrations to learn the **qtl** package.

Reviewer:

Qing Zhou
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095, United States of America
E-mail: zhou@stat.ucla.edu
URL: http://www.stat.ucla.edu/~zhou/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 32, Book Review 5 *Published:* 2010-01-19
January 2010